Abstract:

In this talk, I'll provide an extensive introduction to parallel and distributed machine learning. I'll answer the questions "How actually big is the big data?", "How much training data is enough?", "What do we do if we don't have enough training data?", "What are platform choices for parallel learning?" etc. Over an example of k-means clustering, I'll discuss pros and cons of machine learning in Apache Pig, MPI, DryadLINQ, and CUDA. Time permitting, I'll take a dive into a super large scale text categorization task.

Bio:

Ron Bekkerman is a computer engineer and scientist whose experience spans across disciplines from video processing to business intelligence. Currently a senior research scientist at LinkedIn, he previously worked for a number of major companies including Hewlett-Packard and Motorola. Ron completed his PhD in Computer Science at the University of Massachusetts Amherst in 2007. He holds BSc and MSc degrees from the Technion---Israel Institute of Technology. Ron's research interests lie primarily in the area of large-scale unsupervised learning. He is the corresponding author of several publications in top-tier venues, such as ICML, KDD, SIGIR, WWW, IJCAI, CVPR, EMNLP and JMLR.

Speaker:

Ron Bekkerman
Sr. Research Scientist, LinkedIn

Date: Thur., Feb. 23, 2012
Time: 1:30pm
Location: 1305 Newell-Simon Hall